Protein Science (1998), 7:233–242. Cambridge University Press. Printed in the USA.

Domain assignment for protein structures using a consensus approach: Characterization and analysis

SUSAN JONES,[1] MICHAEL STEWART,[1] ALEX MICHIE,[1] MARK B. SWINDELLS,[3] CHRISTINE ORENGO,[1] AND JANET M. THORNTON[1,2]

[1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London, WC1E 6BT, United Kingdom
[2]Department of Crystallography, Birkbeck College, Malet Street, London, WC1E 7HX, United Kingdom
[3]Helix Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba 292, Japan

(RECEIVED July 23, 1997; ACCEPTED September 17, 1997)

Abstract

A consensus approach for the assignment of structural domains in proteins is presented. The approach combines a number of previously published algorithms, and takes advantage of the elevated accuracy obtained when assignments from the individual algorithms are in agreement. The consensus approach is tested on a data set of 55 protein chains, for which domain assignments from four automated methods were known, and for which crystallographers assignments had been reported in the literature. Accuracy was found to increase in this test from 72% using individual algorithms to 100% when all four methods were in agreement. However a consensus prediction using all four methods was only possible for 52% of the dataset. The consensus approach (using three publicly available domain assignment algorithms (PUU, DETECTIVE, DOMAK)) was then used to make domain assignments for a data set of 787 protein chains from the Protein Data Bank. Analysis of the assignments showed 55.7% of assignments could be made automatically, and of these, 13.5% were multi-domain proteins. Of the remaining 44.3% that could not be assigned by the consensus procedure 90.4% had their domain boundaries assigned correctly by at least one of the algorithms. Once identified, these domains were analyzed for trends in their size and secondary structure class. In addition, the discontinuity of each domain along the protein chain was considered.

Keywords: consensus approach; protein structure; structural domain assignment; structural domain database

Structural domains in proteins have been defined as compact, local, semi-independent units (Richardson, 1981). These units comprise sequential or non-sequential parts of the polypeptide chain. With more than 6000 structures currently in the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977) (and the number rising exponentially) an automated method for the identification of such domains is essential if structural domain databases, such as CATH (Orengo et al., 1997), are to be maintained efficiently. To date there have been a number of methods devised for automatic domain assignment. Some early methods were restricted to the identification of continuous domains (e.g., Go, 1983; Zehfus & Rose, 1986), but recently automatic algorithms have been devised that identify discontinuous domains (Holm & Sander, 1994; Zehfus, 1994; Sowdhamini & Blundell, 1995; Siddiqui & Barton, 1995; Swindells, 1995b). Although these individual algorithms work well and can successfully predict with more than 70% accuracy, this is not sufficient for the mass processing of protein structures for inclusion in classifications such as CATH. We have therefore adopted a consensus approach that uses our observation that the accuracy of domain assignments increases substantially with the number of methods that are in agreement.

The four domain assignment methods assessed are PUU (parser for protein unfolding units) (Holm & Sander, 1994), DETECTIVE (Swindells, 1995b), and DOMAK (Siddiqui & Barton, 1995), and a method by Islam et al. (1995). The algorithms by Holm and Sander (1994), Siddiqui and Barton (1995), and Islam et al. (1995) are based upon the premise that a domain will make more internal contacts (i.e., intra-domain contacts) than external contacts (contact with residues in the remainder of the structure). The PUU program (Holm & Sander, 1994) incorporates a harmonic model used to approximate inter-domain dynamics. This algorithm is used to define domains in the FSSP database (Holm & Sander, 1994). The DOMAK algorithm (Siddiqui & Barton, 1995) calculates a "split value" from the number of each type of contact when the protein is divided arbitrarily into two parts. This split value is large when the two parts of the structure are distinct. The DOMAK algorithm formed the basis of the original domain assignments for

Reprint requests to: Susan Jones, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London, WC1E 6BT, United Kingdom; e-mail: sue@bsm.biochem.ucl.ac.uk.

3 Dee, a database of protein domain definitions (URL http://speed.biop.ox.ac.uk:8080/3Dee). However, all of the assignments in this database are now checked by eye and with reference to the literature (G.J. Barton, pers. obs.). The algorithm of Islam et al. (1995) is based on dividing the chain to minimize the density of inter-domain contacts. The DETECTIVE algorithm (Swindells, 1995b) is, in theory, slightly different, being based on the concept that each domain has an identifiable hydrophobic core (Swindells, 1995b). However, it, too, uses intra-molecular contacts in its calculation. All four algorithms can identify both continuous and discontinuous structural domains.

Using the increased accuracy obtained through a consensus approach, we can divide protein structures into those that can be processed with high accuracy using the automated consensus procedure, and those that must be further analyzed by eye. In this article combinations of four independent algorithms are assessed using a small data set of 55 proteins. This was done by using data collated from the extensive results tables cited in the literature for DOMAK, PUU, and the algorithm of Islam et al. (1995) and by using the results produced by the DETECTIVE algorithm. Then three of the domain assignment algorithms (generously made available by the authors) were applied to 787 representative protein chains from the PDB.

Results

For the purposes of this work, a "correct" assignment is one that has the same number of domains and an overlap score of ≥85% when compared with the crystallographers assignment (collated from the literature by Islam et al., 1995) (see Methods).

Evaluating a consensus approach on 55 protein chains

The accuracy of four independently developed algorithms was assessed using a test set of 55 protein chains. When each method was assessed individually, overall accuracy was found to vary between 67 and 76% (mean 72.2%), depending on the algorithm. None of the algorithms achieved the correct assignment of domains for all 55 chains (Fig. 1A). When the assignments were subsequently analyzed in terms of how many domains each chain had (Fig. 1B), it was found that single-domain proteins were predicted with the highest degree of accuracy (mean 85%). Assignments for two-domain proteins also achieved a high degree of accuracy (mean 72%), but it was clear that as the number of domains increased, the problems of separating them into constituent domains became more complex, and the automated approaches became less reliable.

Agreement between combinations of algorithms was found to increase the accuracy of assignments substantially. With a combination of two algorithms, the mean percentage of comparable and correct assignments was 94.3% (Fig. 2A). With a combination of three algorithms this rose to 97.5% (Fig. 2B), finally reaching 100% with a combination of all four algorithms (Fig. 2C). However, as the accuracy of assignments increases with the number of algorithms generating comparable assignments, so the percentage of structures that can be correctly assigned falls. When two algorithms were combined, comparable assignments were only made for, on average, 61.2% of the structures. When three algorithms were combined, this value falls to 55%, and when four algorithms were used, 52%. Hence, the greater the number of algorithms used, the greater the reliability of the domain assignments, but the smaller the number of structures for which automatic domain assignments can be made.

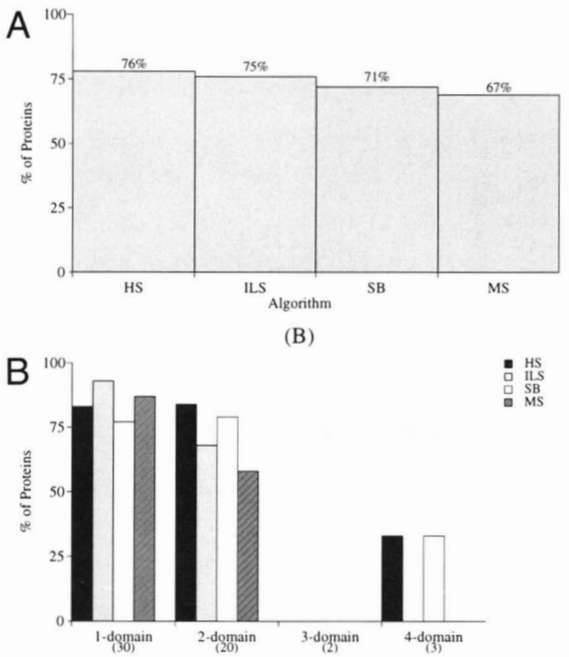

Fig. 1. Accuracy of automatic domain assignment algorithms. **A:** Percentage of proteins given correct domain assignments by four individual algorithms [HS = PUU (Holm & Sander, 1995), ILS = algorithm by Islam et al. (1995), SB = DOMAK (Siddiqui & Barton, 1995), and MS = DETECTIVE (Swindells, 1995b)]. **B:** Percentage of proteins given correct domain assignments by four individual algorithms by domain size. An assignment was defined as correct if they were comparable to those made by crystallographers reported in the literature (collated by Islam et al., 1995).

Applying the consensus approach to the PDB

The increased accuracy obtained by a consensus approach is important for mass processing of structural data from a primary source such as the PDB. The accuracy of domain assignments is critical for us as the assignments form the basis for our domain structure classification CATH (Orengo et al., 1997). At the time of this work, three domain assignment algorithms [DETECTIVE (Swindells, 1995b), PUU (Holm & Sander, 1994), and DOMAK (Siddiqui & Barton, 1995)] were publicly available. On the basis of our previous assessment we assumed that when all three methods were in agreement, the assignment accuracy would be more than 97%. Thus, the three assignment programs were run, and if a consensus was reached, assignments were made automatically. In cases where a consensus was not reached, the chains were assigned manually.

Using the consensus procedure 787 representative proteins from the PDB were assigned to one of three sets: single domain, multi-domain, and unassigned. The breakdown of the assignments made is shown after the consensus procedure (Fig. 3A), and after manual assignments had been made (Fig. 3B). The consensus procedure enabled the assignment of domains for 438 (55.7%) of protein structures automatically, and of these 379 (86.5%) were single-domain proteins and 59 (13.5%) were multi-domain. Of the 349 proteins that were unassigned at the end of the automated proce-

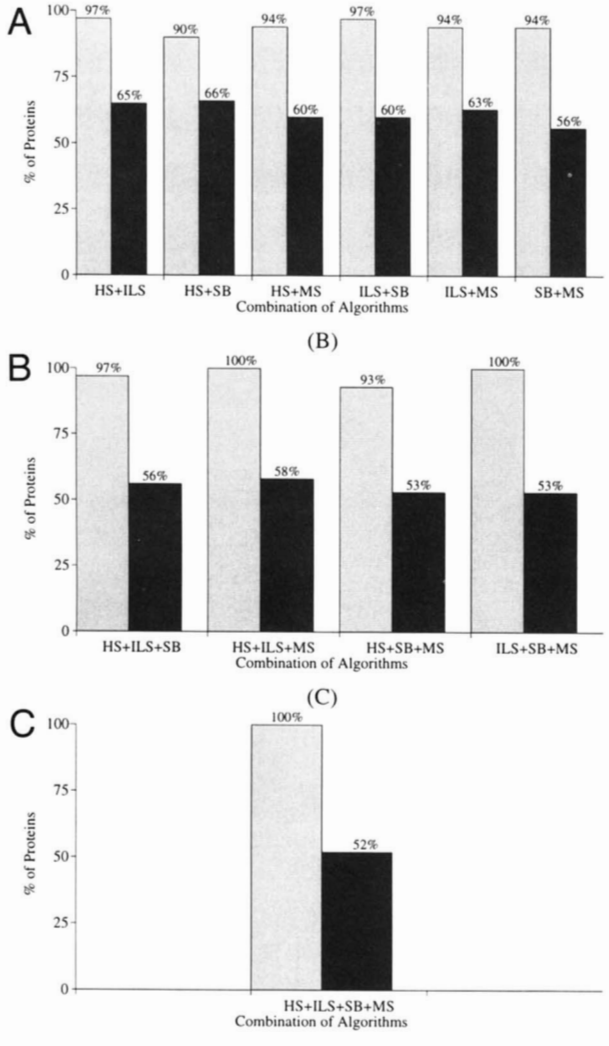

Fig. 2. Analysis of the consensus approach to domain assignments. The percentage of proteins (the number of protein structures as a percentage of the number of protein chains in the test data set (55 chains)) given correct domain assignments by (**A**) combinations of two algorithms (**B**) combinations of three algorithms (**C**), combinations of four algorithms. Where combinations of algorithms have been used data are shown for the percentage of proteins where the combination of algorithms agree (light bars) and the percent percentage of structures that can be correctly assigned domains (dark bars). The initials used to identify the algorithms are shown in the legend to Figure 1.

dure 92 were structures for which one or more of the algorithms failed to make any assignment. Some of these non-assignments were caused by inconsistencies in secondary structure assignment files that were required by all the algorithms. These 349 structures were assigned manually, and this gave 143 single-domain and 206 multi-domain structures.

To validate the consensus method on this scale all those proteins assigned automatically (438) were also checked manually, and of these, only 9 assignments were altered. Three structures assigned as single-domain proteins were re-assigned as multi-domain structures; four structures assigned as two-domain proteins were re-assigned as single-domain structures; and two proteins automatically assigned two domains were re-assigned as three-domain proteins. This gives a 89.8% accuracy for those multi-domain assignments made by the automatic part of the consensus procedure, and a 98.8% accuracy for the single-domain assignments. When considering all of the representative protein structures for which assignments were made, at least one algorithm gave a correct assignment (or one that only required minor adjustments) for 90.4% of the protein structures.

Comparing the final domain assignments (made after the consensus and manual procedures) with the assignments made by each individual algorithm automatically, the accuracy for any one algorithm was, on average, 76.3%. For single-domain proteins the accuracy for any one algorithm was, on average, 87.3%, and for multi-domain proteins, 53.3%. It was observed that there was little difference in the performance of the three algorithms, as all three operated well on some structures but gave poor assignments for others. In general, when PUU and DOMAK made a wrong assignment, they divided a structure into too many domains rather than too few, while the DETECTIVE algorithm was more likely to divide a structure into too few domains.

A recurrent disagreement between algorithms is seen in the TIM Barrel structures. Dependent upon the specific protein in question, the algorithms variously divided these structures into one, two, and three domains, examples of which can be seen in Figure 4. Failure to recognize known folds was a problem that was observed many times, and a further example is shown in Figure 5A. The neuraminidase structure (Burmeister et al., 1992) is a six-bladed β-propellor fold, where all six blades comprise a single domain. Two algorithms assign more than one domain, defining two or three blades to a single domain. One problem with assigning such proteins is that they do not fit the optimized parameters commonly used by the algorithms as well as other domains. In particular, they are larger than the average domain size and many of the algorithms thus try to divide chains of this size wherever possible. One way of improving this situation would be to use information about the folds that are currently known to exist. None of the algorithms used here take advantage of the sets of folds currently known, and this is clearly an area where improvements could be attained. One such algorithm has been described (Holm & Sander, 1996), and another is known to be in development (W.R. Taylor, pers. obs.). However, no detailed reports of their efficacy have yet been published.

The additional examples in Figure 5B–C emphasize the difficulty and subjective nature of domain assignment for protein structures. For adenylosuccinate synthetase (Poland et al., 1993) (Fig. 5B) the three-domain assignment was accepted, as this divided the structure into an α (non-bundle) domain, a three-layer $\alpha\beta$ domain, and a third domain that was classified in CATH as an $\alpha\beta$ complex domain (Orengo et al., 1997). The two-domain assignment produced two rather large domains that would both be classified as $\alpha\beta$ complex. The assignment of five domains clearly divides the protein into too many domains, assigning two very small domains that only have two or three secondary structures. Dihydrolipoamide dehydrogenase (Mattevi et al., 1993) contains 477 residues and has been variously divided into one, two, and four domains (Fig. 5C). For the CATH classifications (Orengo et al., 1997) the structure was actually divided into three domains by manual inspection [Fig. 5C(iv)], to give one two-layer $\alpha\beta$ domain, and two three-layer $\alpha\beta$ domains.

Analysis of domain characteristics

After the whole procedure was completed (including the consensus assignments and those made manually), the division between sin-

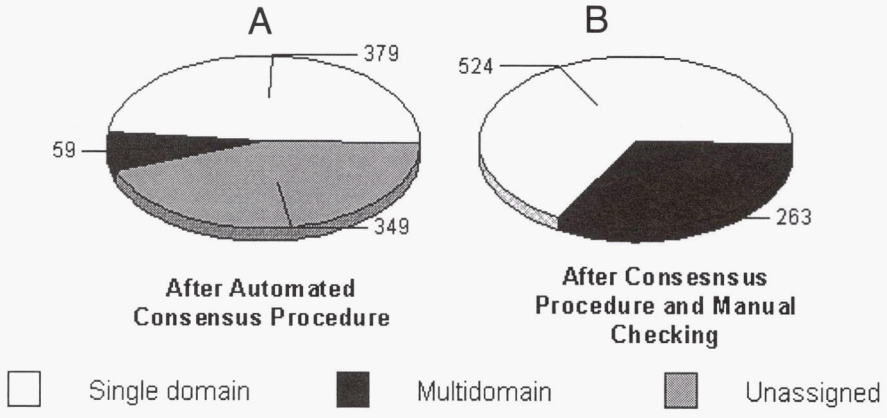

Fig. 3. The percentage of proteins assigned to single and multi-domain categories by the consensus (**A**) and manual procedures (**B**) in the domain assignment procedure. The results are shown for a data set of 787 35% sequence representatives from the PDB.

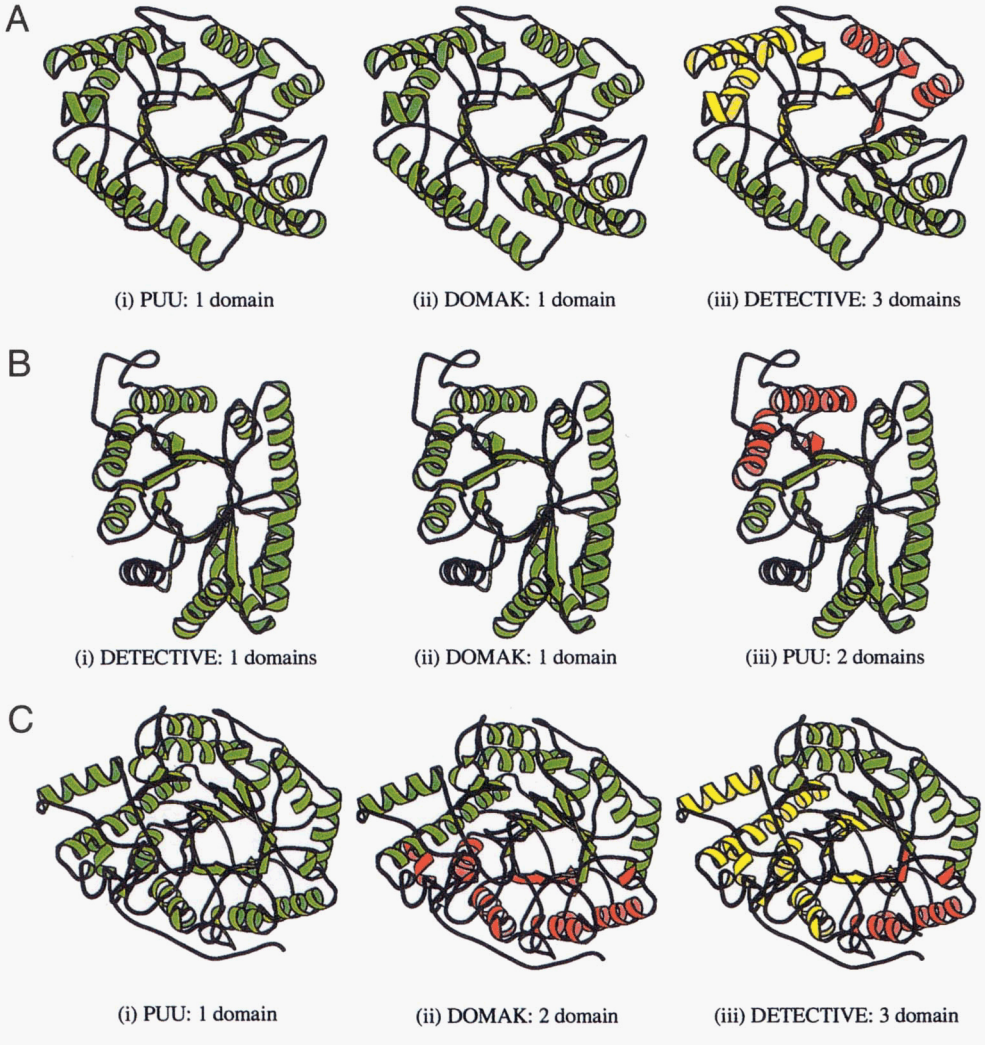

Fig. 4. MOLSCRIPT (Kraulis, 1991) diagrams showing the automatic domain assignments from three algorithms for TIM barrel structures. **A:** 1,4-Beta-D-xylan-xylanohydrolase (1xyz) (Dominguez et al., 1995). **B:** Bacterial luciferase (1brl, chain B) (Fisher et al., 1995). **C:** Beta amylase (1btc) (Mikami et al., 1992). For each structure the assignments made by each algorithm are indicated, with each domain a different color.

Fig. 5. MOLSCRIPT (Kraulis, 1991) diagrams showing the automatic domain assignments from three algorithms for (**A**) Neuraminidase (1nsc, chain A) (Burmeister et al., 1992). (**B**) Adenylosuccinate synthetase (1ade, chain A) (Poland et al., 1993). (**C**) Dihydrolipoamide dehydrogenase (1lpf, chain A) (Mattevi et al., 1993). For each structure the assignments made by each algorithm [and in 5D(iv) a manual assignment] are indicated, with each domain a different color.

gle and multi-domains was analyzed (Fig. 6A). This distribution is similar to that observed for much smaller non-redundant data sets (Holm & Sander, 1994; Islam et al., 1995; Siddiqui & Barton, 1995; Sowdhamini et al., 1996), with single-domain protein comprising 66.8% of the dataset, and with two-domain proteins by far the most commonly occurring multi-domain structures.

The number of domains assigned to a protein shows an increase with the mean size of the protein (in terms of the total number of residues) (Fig. 6B). However, the range of protein sizes forming any number of domains is very large: for example, the smallest protein divided into two domains has 106 residues [porcine spasmolytic protein (1pcp)] and the largest 824 residues [glycogen phosphorylase (1abb, chain A)]. The outliers on this graph were carefully examined to ensure that an incorrect domain assignment had not been made. Outliers to the right of the graph would indicate that a structure had been divided into too few domains, and, indeed, two incorrectly assigned structures were found in this manner. The outliers that still remain have all been manually assigned.

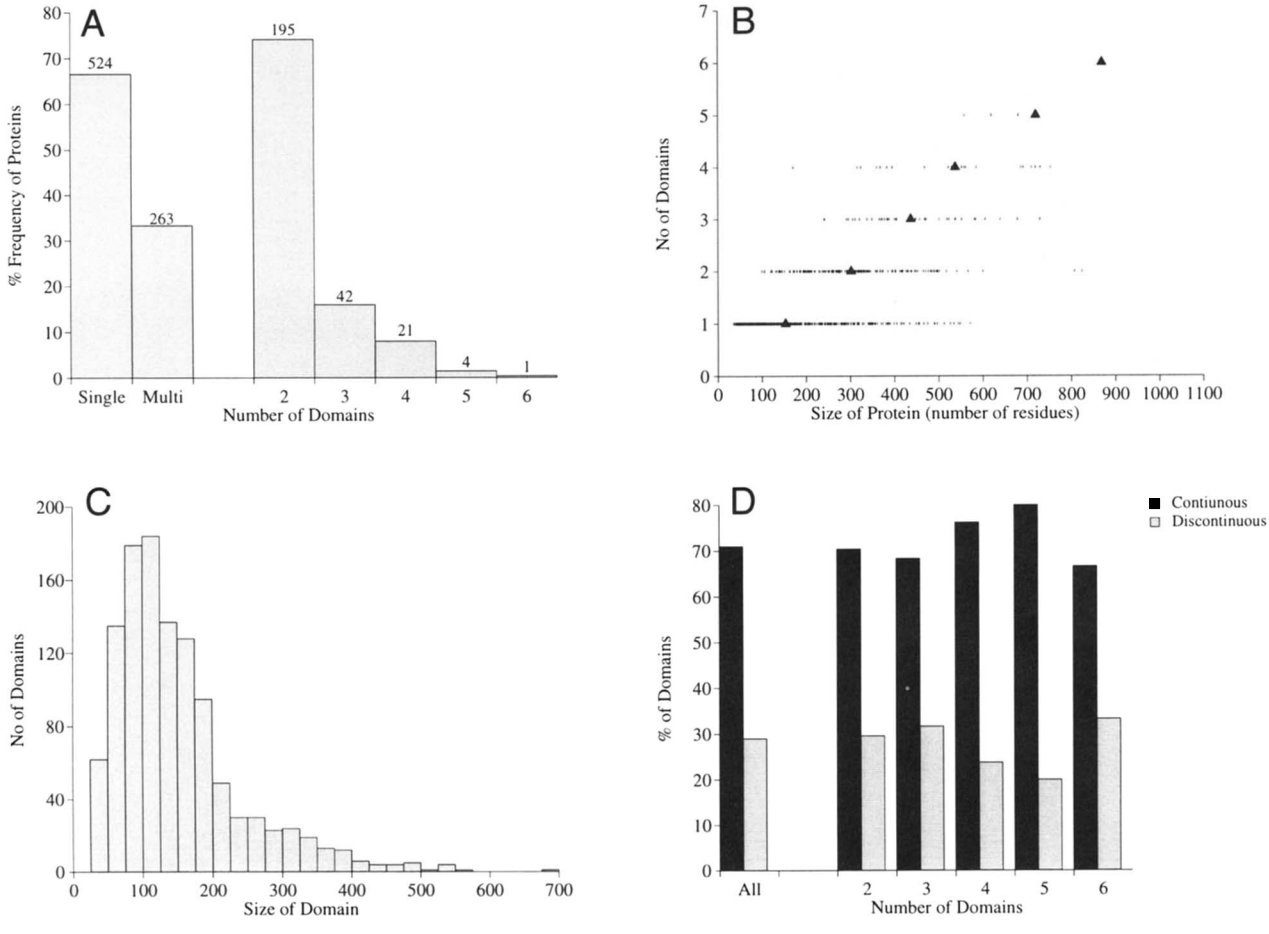

Fig. 6. Analysis of protein domains. Data are shown for a data set of 787 representative protein chains (1146 domains). **A:** Distribution of domains in protein structures. **B:** Protein size and domain assignments. **C:** Distribution of domains sizes. **D:** Discontinuity of domains.

The two outliers on the graph assigned two domains are lipoxygenase-1 (2sbl, chain A) (Boyington et al., 1993), which has 807 residues, and glycogen phosphorylase (1abb, chain A) (Leonidas et al., 1992), which has 824 residues. The lipoxygenase-1 comprises one β sandwich domain and one large mainly α domain that has a complex structure. The glycogen phosphorylase structure comprises one $\alpha\beta$ 3-layer sandwich domain and one large $\alpha\beta$ complex domain.

The size of individual domains also varies widely (Fig. 6C), from 36 residues in E-selectin (1esl) (Graves et al., 1994), a two-domain protein, to 692 residues in lipoxygenase-1 (2sbl, chain B) (Boyington et al, 1993), also a two-domain protein. However, very large domains are the exception. The distribution peaks at around 100 residues per domain and 80.3% of the domains are comprised of less than 200 residues. Very similar distributions have been observed in smaller non-redundant data sets. Siddiqui and Barton (1995), using DOMAK to assign domains for a data set of 230 protein chains, found that 90% of domains comprised less that 200 residues. Holm and Sander (1994) using PUU on a dataset of 330 protein chains, also observed a domain size distribution that peaked at 100 residues.

Domains can be comprised of sequential or non-sequential parts of the polypeptide chain. The occurrence of each type has been listed by the number of domains assigned (Fig. 6D). Sequential domains comprise, on average, 72% of the total, and this applies approximately for all numbers of domains. This figure is similar to that observed by Holm and Sander (1994) who found that 75% of domains they assigned were continuous. Of the total number of multi-domain proteins 44.8% have one or more non-sequential domains. The three algorithms identified proteins with at least one non-sequential domain with comparative results. However, approximately one-third of incorrectly assigned protein structures contained one or more non-sequential domains compared to an 18% occurrence in the whole dataset; giving a clear indication that sequential domains are easier to assign than non-sequential.

The relative sizes of domains in two-domain proteins has been analyzed, by calculating the number of proteins where both domains are of approximately equal size, and where the two domains are of significantly different size (Fig. 7). The results show that over 60% of two-domain proteins are comprised of domains that are of approximately equal size, with each domain contributing between 40–50% of the total number of residues. However, there are some proteins where one domain contributes only 10–20% of the total number of residues; for example, [amylase (1ppi), which comprises one domain with 92 residues (18.4% of the total] and one with 402 residues, which is a $\alpha\beta$ barrel structure.

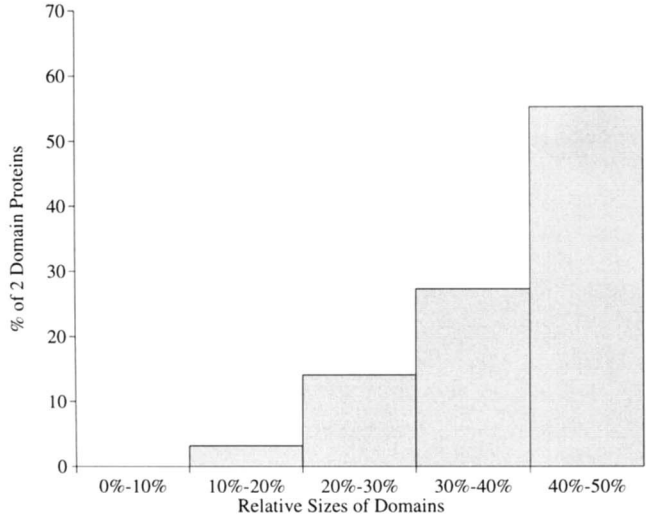

Fig. 7. Comparative size of domains in two-domain proteins. The comparative size is calculated on the basis of the number of residues in each domain and the figure shows the number of two-domain proteins (as a percentage of the total) that have domains comprising 0–10% of the total number of residues in the protein, 10–20%, etc. Hence, those proteins with domains that comprise between 10 and 20% of the total number of residues essentially have one large and one small domain. Conversely, those proteins with domains comprising 40–50% of the total number of residues have domains that are of approximately equal in size.

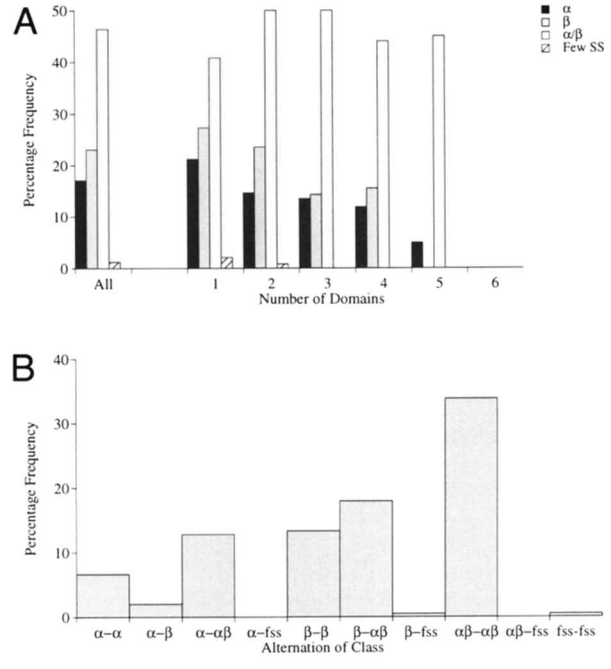

Fig. 8. Domains and secondary structure class (**A**). Percentage frequency of domains (number of domains out of a total of the 1146 domains in the dataset of 787 representative proteins) in each of four structural classes (α, β, $\alpha\beta$, and a few secondary structures). (**B**) Alternation of secondary structural classes in two-domain proteins. The four classes are alpha (α), beta (β), alpha/beta ($\alpha\beta$), and a few secondary structures (fss).

The secondary structure classification of domains into one of four classes, α, β, $\alpha\beta$, and few secondary structures is shown in Figure 8A. $\alpha\beta$ domains are by far the most common, for all numbers of domains, followed by β domains and then α. A similar pattern was observed in both the single-domain proteins and each of the multi-domain proteins. The combination of secondary structure classes is shown for two-domain proteins in Figure 8B. Of the 10 possible combinations, most two-domain proteins comprise two $\alpha\beta$ domains. The other common combinations were one β domain with one β domain, and two β domains.

Discussion

The analysis of the distribution of domain sizes, occurrence of multi-domain proteins and occurrence of sequential domains, gives, for this large non-redundant data set of proteins, similar results to those previously observed in much smaller data sets (Holm & Sander, 1994; Islam et al., 1995; Siddiqui & Barton, 1995; Sowdhamini et al., 1996). Here the size of domains is shown to be an important factor in the assessment of domain assignments, with similar-sized proteins having the same number of domains assigned (although the distribution is wide). This relates to the size of the individual domains that is most commonly around 100 residues. Very large domains are rare, and the presence of extremely large domains (in excess of 800 residues) can, in some cases, indicate an incorrect domain assignment.

The comparison of a series of automatic domain assignment methods highlights the need to consider domain assignments from more than one source. The algorithms operate well on some structures but give poor assignments for others. The DETECTIVE (Swindells, 1995) and PUU (Holm & Sander, 1994) algorithms do particularly well at identifying single-domain proteins. The PUU algorithm also does well identifying some large complex folds that are single domains such as the β-propellors, which can be divided incorrectly into multi-domains by the other algorithms. Both PUU and DOMAK correctly assign many multi-domain structures but can tend to over-divide structures, defining extremely small fragments as domains. It should be emphasized that in our implementation all the algorithms were used as "black boxes," with no manipulation of any parameter settings and no post-processing of the output. DOMAK can provide some post-processing of its domain assignments, using a series of screens (that include the assessment of the discontinuity and size of domains) to indicate the probable agreement with an expected standard assignment (Siddiqui & Barton, 1995). Thus, a more sophisticated use of the algorithms, combined with the lessons learned here, may further elevate the success rate.

The consensus procedure described here has been used to divide protein structures into domains for the CATH classification (Orengo et al., 1997) (URL: http://www.biochem.ucl.ac.uk/bsm/cath). The combination of automatic and manual procedures places CATH between two other protein structure classifications, FSSP (Holm & Sander, 1994) and SCOP (Murzin et al., 1995) in terms of domain assignment. In FSSP domains are assigned totally automatically using the PUU algorithm (Holm & Sander, 1993), and in SCOP (Murzin et al., 1995) all the assignments are made manually by visual inspection. As the size of the PDB grows, at an almost exponential rate, totally automatic procedures for reliable domain assignments have to be the goal.

A consensus approach, combining and comparing a number of algorithms that operate on different criteria, similar to the one described here, is a pragmatic first step for making domain assignments

A (i) Assignments

Residue	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
A	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3
B	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3

(ii) Overlap table:

	A1	A2	A3
B1	6	0	0
B2	1	8	0
B3	0	1	4

(iii) Overlap Score: $\frac{6+8+4}{20} \times 100 = 90\%$

B (i) Assignments

Residue	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
A	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3
B	3	3	1	1	1	1	1	1	1	2	2	2	2	3	3	3	3	3	3	3

(ii) Overlap table:

	A1	A2	A3
B1	5	2	0
B2	0	4	0
B3	2	3	4

(iii) Overlap Score: $\frac{4+5+4}{20} \times 100 = 65\%$

Fig. 9. Examples of overlap score calculations. This score was used to give a quantitative measure of how similar any two assignments were. **A:** Both algorithms assign three domains and the domain boundaries are only extended by a few residues (i), giving an overlap score of [85% (ii, iii)]. These two assignments are comparable. **B:** Both algorithms assign three domains, but the boundaries occur in different residue positions (i), giving an overlap of <85% (ii, iii). These two assignments are not comparable.

for large numbers of protein structures quickly, and with a high degree of confidence. The method described here could be improved by incorporating additional domain assignment algorithms, and incorporating a knowledge based automatic post-processing of the domain assignments, to take into account such parameters as domain size, domain discontinuity, and the structure of known protein folds.

Methods

Evaluating the consensus approach

A dataset of 55 protein chains was chosen to include those for which domain assignments from three of the automated methods (PUU, DOMAK, and the algorithm of Islam et al., 1995), and crystallographers' assignments, had been reported in the literature (Table 1). Domain assignments from the DETECTIVE program were generated for these 55 chains. The four domain assignments made for each of the 55 chains were then compared using a "over-

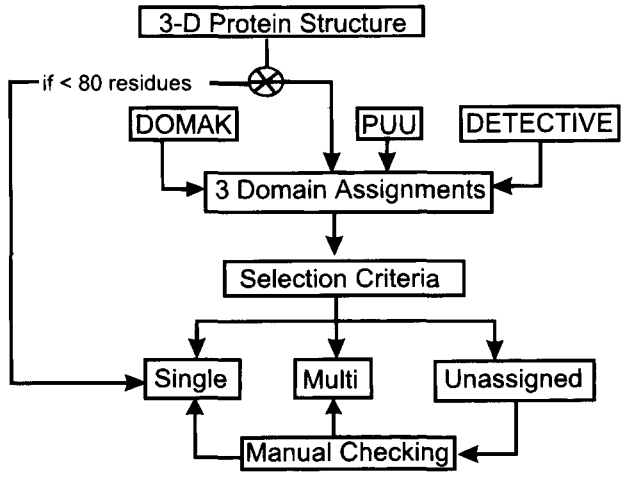

Fig. 10. Flow diagram of the consensus domain assignment procedure.

Table 1. *Dataset of 55 protein chains used to compare four-domain assignment algorithms, with each other and with crystallographer's assignments*[a]

1 Domain			2 Domain		3 Domain	4 Domain
1aak	1ofv	2azaA	1ezm	2cyp	1phh	1atnA
1ace	1pyp	2ccyA	1fnr	2had	3grs	2pmgA
1bbhA	1rbp	2rn2	1gpb	3cd4	8acn	
1bbpA	1rcb	2stv	1lap	3gapA		
1brd	1rveA	2tmvP	1pfkA	3pgk		
1fxiA	1snc	3chy	1ppn	4gcr		
1gky	1tie	3cla	1rhd	5fbp		
1gmfA	1tlk	3dfrA	1sgt	8adh		
1gmpA	1ula	4blmA	1vsg	8atcA		
1gox	1wsyA	5p21	1wsyB	8atcB		

[a]The four-letter code from the PDB is shown for each, with the chain identifier if applicable. Each protein chain has been assigned to a domain category according to the number of domains assigned by the crystallographers, as collated from the literature by Islam et al. (1995).

lap score" similar to that used by Islam et al. (1995). The overlap score was designed to give a quantitative measure of how similar any two assignments were. The method is explained by reference to the examples outlined in Figure 9. For evaluation, any two assignments were considered comparable if they had the same number of domains assigned and an overlap score of ≥85%. This threshold was chosen to take into consideration the omission of small parts of protein structures from some assignments, and also to allow some degree of flexibility in the determination of an acceptable comparison. For the purposes of this work, a "correct" assignment is one that has the same number of domains and an overlap score of ≥85% when compared with the crystallographers assignment (collated from the literature by Islam et al., 1995).

The results (see Results) showed the necessity of using assignments from more than one algorithm, and this has been implemented in a consensus procedure used for the CATH classification (Orengo et al., 1997).

Domain assignments for the PDB and the CATH classification

The consensus approach, as used for assigning domains for protein in the CATH database (Orengo et al., 1997), incorporated three algorithms that were publicly available. These were PUU (Holm & Sander, 1994), DETECTIVE (Swindells, 1995b), and DOMAK (Siddiqui & Barton, 1995). Assignments were made for a data set of 787 35% sequence representative structures (i.e., all proteins with less than 35% sequence identity) taken from the PDB in September 1996. The consensus procedure is summarized in the flow diagram in Figure 10. Any protein with less than 80 residues was automatically assigned as a single-domain structure. All other structures were assigned domains by each of the three domain assignment programs. These assignments were then compared and an assignment made based on a number of selection criteria.

Domain assignments were made automatically (without manual checking) if all three assignments were comparable. As described above, comparable assignments were those for which the number of domains assigned was equal and the overlap score was ≥85%. When an automatic assignment was made the domain boundaries were taken from the DETECTIVE program. For proteins where less than three assignments were comparable the structures were assigned manually, by examining the structures by eye using Rasmol Molecular Renderer (Sayle & Milner-White, 1995). In such cases domain assignments were made by choosing what was determined to be the best assignment made by one of the algorithms, a new assignment, or an alternative assignment obtained from the literature.

After all domain assignments had been made, the 787 representative structures were examined for a number of characteristics including size of domains, segmentation, secondary structure class, and alternation of class.

References

Bernstein F, Koetzle T, Williams G, Meyer G, Meyer E, Brice M, Rodgers K, Kennard O, Shimanouchi T, Tasmui M. 1977. The protein data bank: A computer based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Boyington JC, Gaffney BJ, Amzel LM. 1993. The 3-dimensional structure of an arachidonic-acid 15-lipoxygenase. *Science 260*:1482–1486.

Burmeister WP, Ruigrok RWH, Cusack S. 1992. The 2.2 Å resolution crystal-structure of influenza-B neuraminidase and its complex with sialic acid. *EMBO J 11*:49–56.

Dominguez R, Souchon H, Spinelli S, Dauter Z, Wilson KS, Chauvaux S, Beguin P, Alzari PM. 1995. A common protein fold and similar active site in two distinct families of beta-glycanases. *Nat Struc Biol 2*:569–576.

Fischer AJ, Raushel FM, Baldwin TO, Rayment I. 1995. Three dimensional structure of bacterial luciferase from vibrio harvey I at 2.4 angstroms. *Biochemistry 34*:6581–6586.

Go M. 1983. Modular structural units, exons and function in chicken lysozyme. *Proc Natl Acad Sci USA 80*:1964–1968.

Graves BJ, Crowther RL, Chandran C, Rumberger JM, Li S, Huang KS, Presky DH, Familletti PC, Wolitzky BA, Burns DK. 1994. Insight into E-Selectin/ligand interaction from the crystal structure and mutagenesis of the LEC/EGF domains. *Nature 367*:532–538.

Holm L, Sander C. 1993. Parser for protein folding units. *Proteins Struc Funct Genet 19*:256–268.

Holm L, Sander C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res 22*:3600–3609.

Holm L, Sander C. 1996. Mapping the protein universe. *Science 273*:595–602.

Islam SA, Luo J, Sternberg MJE. 1995. Identification and analysis of domains in proteins. *Protein Eng 8*:513–525.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Mol Graphics 11*:134–138.

Mikami B, Sato M, Shibata T, Hirose M, Aibara S, Katsube Y, Morita Y. 1992. Three-dimensional structure of soybean beta-amylase determined at 3.0 A resolution: Preliminary tracing of the complex with alpha-cyclodextrin. *J Biochem (Tokyo) 112*:541–546.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536–540.

Orengo CA, Michie A, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—A hierarchic classification of protein domain structures. *Structure 5*:1093–1108.

Poland BW, Silva MM, Serra MA, Cho Y, Kim KH, Harris EMS, Honzatko RB. 1993. Crystal structure of adenylosuccinate synthetase from *Escherichia coli*. *J Biol Chem 268*:25334–25342.

Sayle RA, Milner-White EJ. 1995. RASMOL—Biomolecular graphics for all. *Trends Biochem Sci 20*:374–376.

Siddiqui AS, Barton GJ. 1995. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci 4*:872–884.

Sowdhamini R, Blundell TL. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci 4*:506–520.

Sowdhamini R, Rufino SD, Blundell TL. 1996. A database of globular protein structural domains: Clustering of representative family members into similar folds. *Folding Design 1*:209–220.

Swindells MB. 1995a. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci 4*:93–102.

Swindells MB. 1995b. A procedure for detecting structural domains in proteins. *Protein Sci 4*:103–112.

Zehfus MH. 1994. Binary discontinuous compact protein domains. *Protein Eng 7*:335–340.

Zehfus MH, Rose GD. 1986. Compact units in proteins. *Biochemistry 25*:5759–5765.